5.1 Design of the Power6™ Microprocessor

Joshua Friedrich[1], Bradley McCredie[1], Norman James[1], Bill Huott[2], Brian Curran[2], Eric Fluhr[1], Gaurav Mittal[1], Eddie Chan[1], Yuen Chan[2], Donald Plass[2], Sam Chu[1], Hung Le[1], Leo Clark[1], John Ripley[1], Scott Taylor[1], Jack Dilullo[1], Mary Lanzerotti[3]

[1]IBM Systems Group, Austin, TX
[2]IBM Systems Group, Poughkeepsie, NY
[3]IBM Research, Yorktown Heights, NY

POWER6 ™ doubles the operating frequency of POWER5 ™ at constant power, includes a scalable, high-bandwidth memory subsystem capable of supporting 128-way SMP systems, and contains mainframe-like reliability, availability and serviceability. This 341mm^2 dual-core microprocessor shown in Fig. 5.11 is fabricated in a 65nm SOI process with 10 levels of low-dielectric copper interconnects [1] and contains over 700M transistors. It operates at clock frequencies exceeding 5GHz in high-performance applications, and can also operate under 100W in power-sensitive applications.

Simultaneous dual-threaded execution, load lookahead, and enhanced data and instruction prefetch capabilities drive the performance of the in-order superscalar cores. Two hardware accelerators, a VMX unit for multi-media processing and a decimal floating-point unit (DFU), also deliver performance to key customer applications. The memory subsystem includes 4MB of private level-2 cache for each core, an L3 controller that supports 32MB of off-chip level-3 cache, two on-chip asynchronous memory controllers, and an SMP interconnect fabric.

POWER6's mainframe-like reliability begins with extensive error checking and recovery. ECC protects all large, on-chip caches, parity guards more than 99% of small register arrays and more than 70% of the dataflow logic, and hundreds of logical cross-checkers monitor the control logic. If an error is detected, the checkpoint and recovery unit (RU) allows the core to safely restart operation prior to the offending instruction without impacting the system.

To meet the frequency goals of POWER6 while supporting this architectural and RAS complexity, the physical design team employed significant innovations. Other processor designs have increased frequency through excessive pipelining which reduces IPC and exponentially increases area and power for little real performance gain. In contrast, POWER6 maintains the POWER5 pipeline depth for key execution units, while doubling the frequency and performance of these units. For example, POWER6 reduces the POWER5 FXU- and FPU-dependent operation latencies by 70% and 40% respectively [2].

These performance enhancements result from joint high-frequency design by the logic, circuit, integration, and characterization teams. Cross-sections and floor plans for the chip's IPC-critical logic and array paths began in the concept phase, and a new tool for obtaining accurate timing based on floorplanned schematics allows detailed tuning prior to physical implementation. The final designs of these critical paths minimize logic complexity and employ high-speed techniques such as pulsed multiplexer latches and post-layout circuit tuning. The arrays also leverage high speed topologies such as the dynamic NOR pre-decode shown in Fig. 5.1.2, which uses device ratios and pulsed clocks to minimize stacked transistors in the critical path. Cycle-time margin in the memory subsystem and non-critical core paths, graphically illustrated in Fig. 5.1.3, ensures that manufacturing variations on under-designed paths do not artificially limit the frequency of these optimized paths, and also enables a low-cost, core-centric sorting methodology with minimal frequency margins. Empirically altering latch launch and capture times using programmability in the local clock buffers also allows the team to isolate and improve the performance of critical paths in hardware as shown in Fig. 5.1.4.

With system power delivery and cooling capacity relatively fixed, reducing chip power is also critical to delivering high performance. To aid this effort, the POWER6 team developed a "nosim" power tool that quickly and accurately predicts the power of functional blocks without circuit simulation, based on the design's RTL or schematic description. This tool allows careful tracking of power throughout all phases of the design and provides rapid feedback for iterative design improvements. Automated tools also directly help to reduce power by decreasing latch sizes, inserting higher V_t transistors, and reducing FET widths in non-critical timing paths. The design also minimizes AC power by heavily using clock gating and pulsed latches. Between 50% and 85% of all latches are gated at any time, and roughly 80% of the chip's latches operate in a "pulsed" mode, which holds the C1 clock high even when the latch is not gated and converts the C2 clock to a pulse of programmable width. Fig. 5.1.5 illustrates the power distribution within the chip.

In addition to minimizing power consumption, POWER6 also provides ample tools to enable dynamic, system-level power management. Twenty-four digital thermal sensors give detailed feedback of the current on-chip temperature map. Moreover, distributed critical path monitors, designed and empirically calibrated to track with the chip's cycle limiting paths, report the chip's frequency margin at its current operating point. Using these sensor feedbacks, power management routines may modify voltage or frequency to gain performance or reach a lower power state. The chip may also internally regulate power and performance using throttling of instruction fetch or dispatch.

POWER6 operates in low-power systems as well as high-performance servers because it can function below 0.75V and above 1.3V on its main voltage supply. This wide range of voltage operation results from POWER6's robust physical design and its use of three major voltage planes. The power- and performance-sensitive logic supply (V_{DD}) is optimized on a per-chip basis to compensate for process variations and enable optimal frequency within any power constraint for each chip. The array supply (V_{CS}) is set 100 to 200mV higher than V_{DD} to optimize stability and read performance while maintaining low power. The I/O supply (V_{IO}) remains constant across chips enabling sensitive analog circuits in the PLL and I/O to support only a single preferred voltage.

Designing a multi-voltage chip requires significant changes to the design methodology. New tools enable the logic team to specify the voltage domain of functional blocks and signals inside RTL. Tools also validate that the physical implementation match the logical specification and contain correct level-translation circuitry on nets crossing between voltage domains. Initial floorplanning work requires consideration of the split power grid to ensure that each supply reaches the functions it powers. Static timing, global routing, and buffering tools are taught to understand voltage domains and treat signals according to their voltage. Designers also complete extensive SPICE modeling of voltage and process corners to validate functionality across the desired range. Finally, lab characterization identifies circuits that limit the operating voltage minimums (V_{MIN}) or restrict the values of supplies relative to each other (V_{DIFF}). Fig. 5.1.6 shows the expansive V_{DD} and V_{CS} operating range of a typical chip that results from these efforts.

References:

[1] E. Leobandung, H. Nayakama, D. Mocuta et al., "High Performance 65nm SOI Technology with Dual Stress Liner and Low Capacitance SRAM cell", 2005 Symp. VLSI Technology, pp. 126-127, June, 2005.

[2] B. Curran, B. McCredie, L. Sigal, et al., "4GHz+ Low-Latency Fixed-Point and Binary Floating-Point Execution Units for the POWER6 ™ Processor", IEEE J. Solid-State Circuits, pp. 1728 – 1734, Feb., 2006.

- 5+ GHz operation
- >700M transistors
- 341mm^2 die
- 65nm SOI process w/ 10 levels of low-k Cu interconnect
- Two superscalar, simultaneous multi-threaded (SMT) cores with VMX & DFU accelerators
- Core Recovery Unit for error correction
- 8 MB Level-2 cache
- Two on-chip memory controllers & a L3 controller
- 128 way SMP Fabric

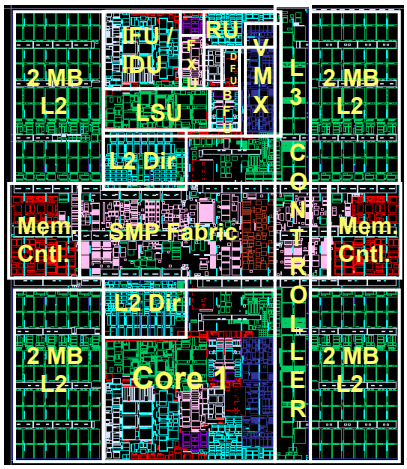

Figure 5.1.1: POWER6 Overview.

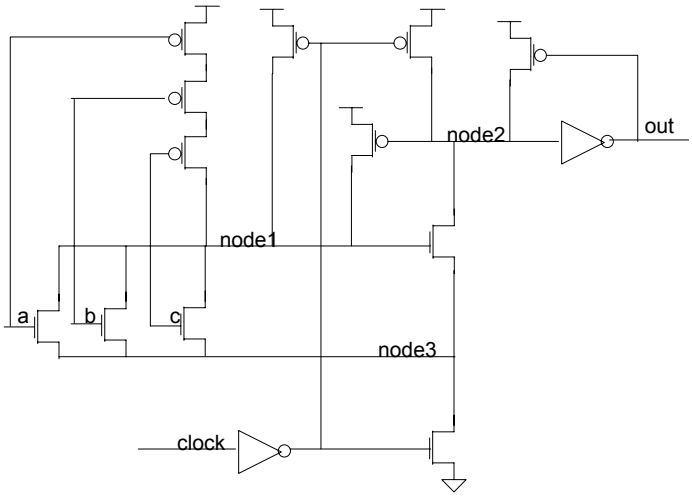

Figure 5.1.2: 3-input Dynamic Address Predecoder.

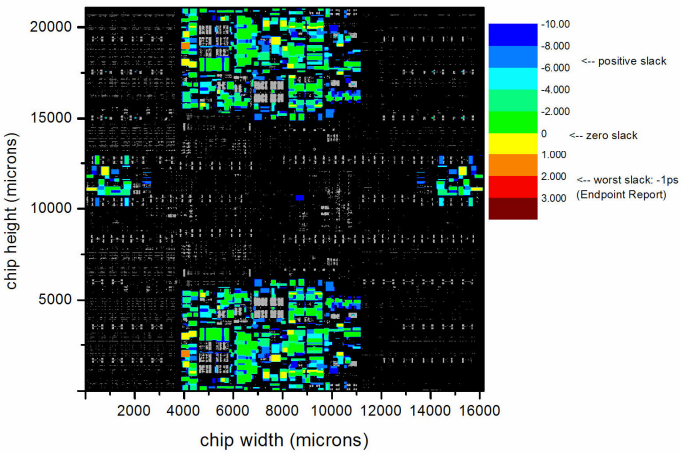

Figure 5.1.3: Location of POWER6 Critical Paths.

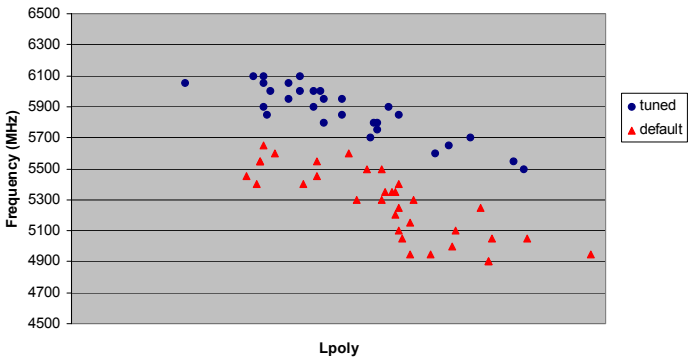

Figure 5.1.4: Frequency Improvement from Hardware Tuning.

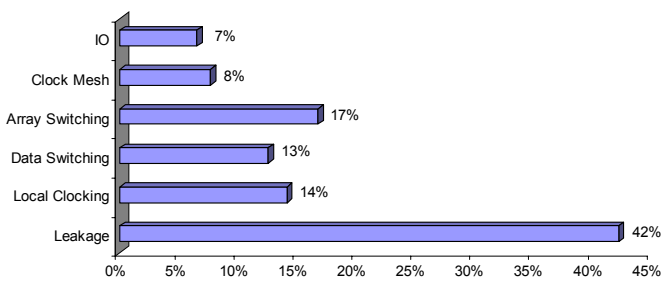

Figure 5.1.5: POWER6 Power Distribution.

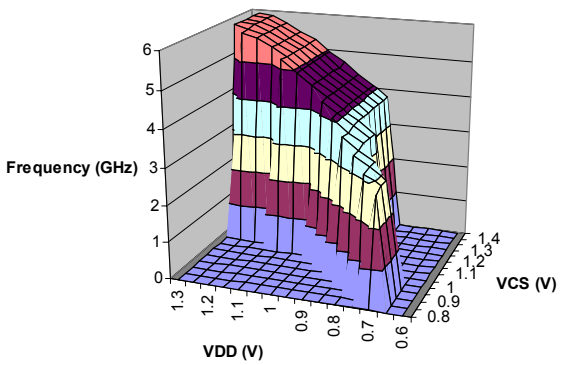

Figure 5.1.6: Fmax vs. VDD and VCS.